[0114] CLAIMS

What is claimed is:

1. A method comprising:

building a data overlay as a data structure on top of a logical space included in a distributed hash table (DHT) for a peer-to-peer system; wherein the logical space includes a plurality of DHT nodes having an associated plurality of DHT zones;

building, in the data overlay, a topology of a tree having a plurality of levels each including one or more tree nodes associated with respective said DHT nodes, wherein:

the first level of the tree includes a single tree node having a single tree node zone corresponding to the entire span of the logical space of the DHT and being logically divided into a plurality of said tree node zones respectively corresponding to:

the tree nodes at each level of the tree; and

parts of the logical space of the DHT;

each said tree node includes a key member which identifies a key associated with its respective tree node zone.

mapping a plurality of machines to the logical space of the DHT, wherein:

each machine corresponds to one or more of more of the tree node zones;

each machine selects as its representative node, from the one or more tree node zones corresponding thereto, the tree node corresponding to the largest size tree node zone; and

each said representative node selects as its parent node another said representative node that is the representative node for an adjacent said tree node zone that has a larger size.

2. The method as defined in Claim 1, further comprising:

gathering metadata at each said machine;

sending the metadata gathered at said machine to the corresponding representative node;

gathering the metadata received by each said representative node; and

sending the metadata gathered by each said representative node to the corresponding parent node; and

gathering metadata received at the single tree node at the first level of the tree.

3. The method as defined in Claim 2, further comprising:

processing the data gathered at the single tree node at the first level of the tree; and

sending the processed data from the single tree node at the first level of the tree to each said machine via the respective parent and representative nodes.

4. The method as defined in Claim 3, wherein:

the metadata comprises information regarding the operation of each said machine; and

the processed data comprises instructions that can govern the operation of each said machine.

5. The method as defined in Claim 1, wherein:

the single tree node zone that corresponds to the entire span of the logical space of the DHT is evenly divided into k tree node zones;

k is the number of tree nodes at the first level of the tree; and

the j-th tree node at level i of the tree has a tree node zone having:

a size of $[j/k^i, (j+1)/k^i]$; and

a key of $(2j+1)/2k^i$; where $(0<=j<2^i)$.

6. The method as defined in Claim 5, wherein:

each said key has a value that is a function of coordinates that identify the center of the respective tree node zone;

the i-th level of the tree contains k^i tree nodes; and

the tree node zone of each tree node has a size of $1/k^i$.

7. The method as defined in Claim 1, further comprising computing, for each said machine, the respective keys of the respective representative and parent nodes for the machine.

8. The method as defined in Claim 7, wherein the computing the respective keys further comprises obtaining information, with the machine, using a look up in the DHT, wherein the machine uses the information with the key of the corresponding said representative node to establish communication with the machine corresponding to the representative node.

9. The method as defined in Claim 1, further comprising:

receiving, at each said machine, a heartbeat transmission from each said machine in an adjacent said tree node zone; and

when any said heartbeat transmission is not timely received, accounting for the absence of the corresponding said machine in the adjacent said tree node zone by:

repeating the providing of the DHT;

repeating the building of the data overlay as the data structure on top of the logical space of the DHT;

repeating the building of the multilevel tree in the rebuilt data overlay; and

repeating the mapping of the plurality of machines to the logical space of the DHT.

10. The method as defined in Claim 1, wherein each said representative node and each said parent node is selected as an optimization function of availability of resources.

11. The method as defined in Claim 10, wherein the optimization function is based upon criteria selected from the group consisting of network coordinates, bandwidth bottleneck, maximal latency, and variance of latencies, whereby the most resource hungry task is performed by the most resource available machines in the peer-to-peer system.

12. The method as defined in Claim 1, wherein:

the DHT governs the insertion and retrieval of objects into and from the peer-to-peer system; and

the logical space includes a plurality of DHT nodes having an associated plurality of DHT zones; and

the data overlay of the DHT is built by:

associating objects in the data structure with the DHT nodes; and

establishing links between the objects in the data structure.

13. .The method according to claim 12, wherein each link includes:

a first field that provides a hardwired pointer that points from a first object to a second object; and

a second field that provides a soft-state pointer that points from the first object to a DHT node which hosts the second object.

14. The method according to claim 12, wherein the building of the data overlay makes use of:

a first primitive for setting a reference that establishes a pointer to an object in the DHT;

a second primitive for returning an object referenced by a pointer; and

a third primitive for deleting an object referenced by a pointer.

15. The method according to claim 1, wherein each tree node in the data overlay includes an operation member which defines an operation that is to be performed on data that is passed through the tree node.

16. The method according to claim 1, wherein each tree node in the data overlay includes a report member which defines a report type that is to be generated using the tree node.

17. The method according to claim 1, wherein:

the first level of the tree includes the tree node that is a root node for the tree; and

the root node corresponds to the tree node zone that corresponds to the entire span of the logical space of the DHT.

18. A computer readable store including machine readable instructions for implementing the building of objects in the data overlay according to the method of claim 12.

19. A computer readable store having stored thereon a data overlay produced according to the method of claim 1.

20. A computer readable store having stored thereon a data structure that comprises a data overlay as a data structure on top of a logical space included in a DHT for a peer-to-peer system; wherein:

the DHT governs the insertion and retrieval of objects into and from a peer-to-peer system;

the logical space includes a plurality of DHT nodes having an associated plurality of DHT zones;

the data overlay of the DHT is built by:

associating objects in the data structure with the DHT nodes; and

establishing links between the objects in the data structure;

the data overlay has a topology of a tree that includes a plurality of levels;

the tree includes a plurality of tree nodes associated with respective said DHT nodes; the tree nodes include a root node having a tree node zone corresponding to the logical space of the DHT;

the tree node zone of the root node is logically divided into a plurality of tree node zones respectively corresponding to:

the number of tree nodes at each level of the tree; and

a part of the logical space of the distributed hash table;

each said tree node includes a key member which identifies a key associated with its respective tree node zone;

the logical space of the DHT is mapped to a plurality of machines;

each machine corresponds to one or more of more of the tree node zones;

each machine selects as its representative node, from the one or more tree node zones corresponding thereto, the tree node corresponding to the largest size tree node zone; and

each said representative node selects as its parent node another said representative node that is the representative node for an adjacent said tree node zone that has a larger size.

21. The computer readable store as defined in Claim 20, wherein:

the tree node zone of the root node is evenly divided into k tree node zones, where k is the number of tree nodes at the first level of the tree; and

the j-th tree node at level i of the tree has a tree node zone having:

a size of $[j/k^i, (j+1)/k^i]$; and

a key of $(2j+1)/2k^i$; where $(0<=j<2^i)$.

22. The computer readable store as defined in Claim 21, wherein:

each said key has a value that is a function of coordinates that identify the center of the respective tree node zone;

the i-th level of the tree contains k^i tree nodes; and

the tree node zone of each tree node has a size of $1/k^i$.

23. The computer readable store as defined in Claim 20, wherein:

the DHT governs the insertion and retrieval of objects into and from the peer-to-peer system;

the logical space includes a plurality of DHT nodes having an associated plurality of DHT zones; and

the data overlay of the DHT:

has objects in the data structure associated with the DHT nodes; and

has links established between the objects in the data structure.

24. The computer readable store as defined in Claim 20, wherein each link includes:

a first field that provides a hardwired pointer that points from a first object to a second object; and

a second field that provides a soft-state pointer that points from the first object to a DHT node which hosts the second object.

25. The computer readable store as defined in Claim 20, wherein:

a first primitive sets a reference that establishes a pointer to an object in the DHT;

a second primitive returns an object referenced by a pointer; and

a third primitive deletes an object referenced by a pointer.

26. The computer readable store as defined in Claim 20, wherein each tree node in the data overlay includes an operation member which defines an operation that can be performed on data that is passed through the tree node.

27. The computer readable store as defined in Claim 20, wherein each tree node in the data overlay includes a report member which defines a report type that is to be generated using the tree node.

28. The computer readable store as defined in Claim 20, wherein:

the first level of the tree includes the tree node that is a root node for the tree; and

the root node corresponds to the tree node zone that corresponds to the entire span of the logical space of the DHT.

29. A peer-to-peer system including a plurality of machines interacting in peer-to-peer fashion, comprising:

a logical space of a DHT that includes a plurality of DHT nodes having a plurality of associated DHT zones, wherein the DHT governs the insertion and retrieval of objects into and from the peer-to-peer system;

a data overlay as a data structure on top of the logical space of the DHT, wherein:

the data overlay of the DHT:

has objects in the data structure associated with the DHT nodes; and

has links established between the objects in the data structure;

the data overlay has a topology of a tree that includes a plurality of levels and includes a plurality of tree nodes associated with respective said DHT nodes;

the tree nodes include a root node having a tree node zone corresponding to the logical space of the DHT;

the tree node zone of the root node is logically divided into a plurality of tree node zones respectively corresponding to:

the number of tree nodes at each level of the tree; and

a part of the logical space of the distributed hash table;

each said tree node includes a key member which identifies a key associated with its respective tree node zone;

the logical space of the DHT is mapped to a plurality of machines;

each machine corresponds to one or more of more of the tree node zones;

each machine selects as its representative node, from the one or more tree node zones corresponding thereto, the tree node corresponding to the largest size tree node zone; and

each said representative node selects as its parent node another said representative node that is the representative node for an adjacent said tree node zone that has a larger size.

30. The system according to claim 29, further comprising routing logic configured to route data through the data overlay by passing the data through the tree nodes.

31. The system according to claim 30, wherein the routing logic is configured to route the data through the data overlay by gathering data from DHT nodes and passing the data up through the tree nodes to the root node of the tree.

32. The system according to claim 30, wherein the routing logic is configured to route data through the data overlay by disseminating data from the root node of the tree, through the tree nodes, to the DHT nodes.

33. An apparatus for building a peer-to-peer system, the apparatus comprising:

means for building a data overlay as a data structure on top of a logical space included in a distributed hash table (DHT) for a peer-to-peer system; wherein:

the DHT governs the insertion and retrieval of objects into and from a peer-to-peer system;

the logical space includes a plurality of DHT nodes having an associated plurality of DHT zones; and

the data overlay of the DHT is built by:

associating objects in the data structure with the DHT nodes; and

establishing links between the objects in the data structure;

means for building a topology of a tree in the data overlay, the tree having a plurality of levels and including a plurality of tree nodes associated with respective said DHT nodes, wherein:

the tree nodes include a root node having a tree node zone corresponding to the logical space of the DHT;

the tree node zone of the root node is logically divided into a plurality of tree node zones respectively corresponding to:

the number of tree nodes at each level of the tree; and

a part of the logical space of the distributed hash table;

each said tree node includes a key member which identifies a key associated with its respective tree node zone;

means for mapping a plurality of machines to the logical space of the DHT, wherein each machine corresponds to one or more of more of the tree node zones;

means for selecting as its representative node, from the one or more tree node zones corresponding to a respective said machine, the tree node corresponding to the largest size tree node zone; and

means for selecting for each said representative node as its parent node another said representative node that is the representative node for an adjacent said tree node zone that has a larger size.

34. The apparatus as defined in claim 33, further comprising:

means for gathering metadata at each said machine;

means for sending the metadata gathered at said machine to the corresponding representative node;

means for gathering the metadata received by each said representative node; and means for sending the metadata gathered by each said representative node to the corresponding parent node; and

means for gathering metadata received at the single tree node at the first level of the tree.

35. The apparatus as defined in claim 34, further comprising:

means for processing the data gathered at the single tree node at the first level of the tree; and

means for sending the processed data from the single tree node at the first level of the tree to each said machine via the respective parent and representative nodes.

36. The apparatus as defined in Claim 35, wherein:

the metadata comprises information regarding the operation of each said machine; and

the processed data comprises instructions that can govern the operation of each said machine.

37. The apparatus as defined in claim 33, further comprising:

means for receiving at a machine a heartbeat transmission from each said machine in an adjacent said tree node zone; and

means, when any said heartbeat transmission is not timely received, for accounting for the absence of the corresponding said machine in the adjacent said tree node zone by:

repeating the providing of the DHT;

repeating the building of the data overlay as the data structure on top of the logical space of the DHT; and

repeating the building of the multilevel tree in the rebuilt data overlay.

38. The apparatus as defined in claim 37, wherein:

the means for accounting further comprises means for repeating the mapping of the plurality of machines to the logical space of the DHT; and

the apparatus further comprises means for selecting each said representative node and each said parent node as an optimization function of availability of resources of the corresponding machines.

39. The apparatus as defined in claim 38, wherein the optimization function is based upon criteria selected from the group consisting of network coordinates, bandwidth bottleneck, maximal latency, and variance of latencies, whereby the most resource hungry task is performed by the most resource available machines in the peer-to-peer system.

40. The apparatus as defined in claim 33, further comprising means for routing data through the data overlay by passing the data through the tree nodes.

41. The apparatus as defined in claim 40, wherein the routing means includes means for routing the data through the data overlay by gathering data from DHT nodes and passing the data up through the tree nodes to the root node of the tree.

42. The apparatus as defined in claim 40, wherein the routing means includes means for routing data through the data overlay by disseminating data from the root node of the tree, through the tree nodes, to the DHT nodes.